Problem Set

Deadline: Monday, Oct. 31, at 11:59pm.

Submission: You need to submit two files through Quercus:

- Your answers to the questions as a PDF file `problem_set_sol.pdf` (We encourage using LaTeX).
- Your implementation of question 3 as a Python file `code.py`.

Make sure your solutions and code structure are readable. You may lose up to 10 pts if we have a hard time reading your write-up/code.

1 Structure identifiability [20 pts]

Discovering causal graphs from observational distributions is generally impossible; we can identify the causal graph up to the skeleton and v-structures (see the observational equivalent theorem in lecture 2). For example, in the case of two dependent variables $X \not\perp\!\!\!\perp Y$, the causal graphs in Figure 1 are all feasible.

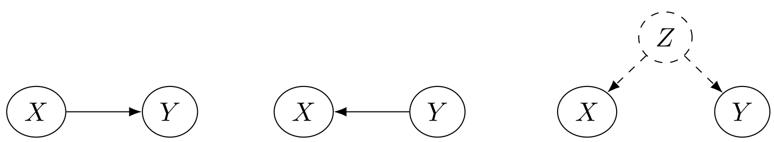


Figure 1

One approach to distinguish cause and effect is to make parametric assumptions on $P_{\text{effect}|\text{cause}}$ and P_{cause}. For some of those parametric classes, there are theoretical results showing that there is only one direction ($P_{Y|X}$ or $P_{X|Y}$) that can satisfy the assumption. In this question, we will learn one such result for **linear** SCMs. Consider the following SCM, which induces joint distribution $P(X, Y)$:

$$Y := \alpha X + \epsilon_Y \quad X \perp\!\!\!\perp \epsilon_Y \tag{1.1}$$

for continuous 1-dimensional random variables Y, X, ϵ_Y.

1. **[5 pts]** Assume X and ϵ_Y are Gaussian. Then, find $\beta \in \mathbb{R}$ such that the following SCM induces the same distribution $P(X, Y)$:

$$X := \beta Y + \epsilon_X \quad Y \perp\!\!\!\perp \epsilon_X \tag{1.2}$$

2. **[10 pts]** Now, assume there exists $\beta \in \mathbb{R}$ and random variable ϵ_X such that

$$X := \beta Y + \epsilon_X \quad Y \perp\!\!\!\perp \epsilon_X \tag{1.3}$$

generates the same distribution $P(X, Y)$. Show that X and ϵ_Y are Gaussian.

Hint: You can use the following theorem

Theorem 1. *Let $X_1, X_2, \ldots, X_n$ be n independent random variables, let*

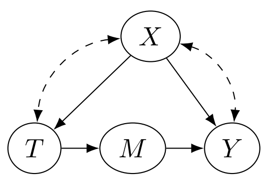
$$\begin{cases} Y_1 = \sum_{i=1}^{n} a_i X_i \\ Y_2 = \sum_{i=1}^{n} b_i X_i \end{cases} \tag{1.4}$$

and suppose Y_1 and Y_2 are independent. Now, if $a_i b_i \neq 0$, then X_i must be Gaussian.

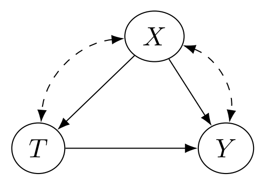
3. **[5 pts]** Suppose we "believe" that the effect is a linear function of cause with non-Gaussian noise. (Informally) describe an algorithm that identifies cause and effect for two variables X and Y.

2 *do*-calculus! [20 pts]

Consider the following causal graphs.



(a)



(b)

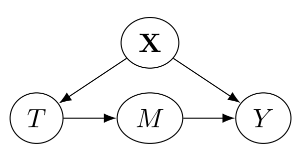
For each graph, is the causal effect of T on Y identifiable? If yes, use *do*-calculus and find a nonparametric formula based on observational distribution. If no, try to prove non-identifiability (e.g., using a counterexample).

3 Parametric estimation of causal effects [40 pts]

For this question, we have provided you with a simulated dataset, `data.csv`, of 5,000 samples with features $\{Y, T, X_0, \ldots, X_9\}$. The goal is to estimate the average treatment effect on Y, i.e.,

$$\text{ATE} = \mathbb{E}[Y|do(T=1)] - \mathbb{E}[Y|do(T=0)] \tag{3.1}$$

Assume the corresponding causal graph as the following ($\mathbf{X} = (X_0, \ldots, X_9)$):



- Since the causal graph satisfies both Front-door and Back-door criteria, try to estimate the ATE using both formulas. In particular, you need to write three Python functions: `frontdoor`, `backdoor_com` (parametric Back-door formula with conditional outcome modeling), and `backdoor_grouped_com` (parametric Back-door formula with grouped conditional outcome modeling).
- Include the result of each of the three methods in your write-up. Also, report one estimation as your final answer. You can choose one of the three outputs, a combination of them, or just a random guess! (or any other causal estimation method). Part of your grading will be based on how close your final answer is to the true ATE.

- Make sure to include your implementation in a file `code.py`. Note that you can use any machine learning library to implement parametric functions for estimating conditional expectations, including linear regression, random forests, neural networks, etc. However, you *cannot* use libraries for causal estimation.

4 Can we identify unobserved confoundings? [20 pts]

Consider an SCM $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathcal{F}, P_{\mathbf{U}})$ with endogenous variables $\mathbf{V} = \{X, \mathbf{T}, Y\}$ and unknown $\mathcal{F}, P_{\mathbf{U}}$. Assume the outcome and covariates are one-dimensional $X, Y \in \mathbb{R}$, and the treatment $\mathbf{T} \in \mathbb{R}^m$. In this question, we will focus on the Back-door causal graph, i.e.,

$$\mathcal{G}_{\mathcal{M}}: \quad \begin{tikzcd} & \boldsymbol{X} \arrow[dl] \arrow[dr] & \\ \boldsymbol{T} \arrow[rr] & & Y \end{tikzcd}$$

We know that in the case of unobserved confoundings X, the average treatment effect

$$\text{ATE} = \mathbb{E}[Y|do(\mathbf{T} = \mathbf{t})] - \mathbb{E}[Y|do(\mathbf{T} = \mathbf{t}')] \tag{4.1}$$

is not identifiable. In lecture 3, we showed that by constructing one example with $m = 1$. The reason for non-identifiability is that multiple choices of $P(Y|X)$ and $P(\mathbf{T}|X)$ can result in the same observed distribution $P(\mathbf{T}, Y)$. But, since we do not observe X, we cannot tell which one is the real one.

In this question, we want to test a hypothesis: *What if we can identify the latent factor?* In particular, we will consider the case of multi-dimensional treatments with $m > 1$. The intuition is that if we observe more causes, where each of them has some information about the unobserved confounding, it will be more likely to identify X. After identifying the confounder, we can substitute it with an estimation of it and use methods such as parametric G-formula to estimate the causal effect.

1. **[5 pts]** Let's formalize the above idea. To identify the hidden confounder, we can fit a probabilistic factor model [1] to capture the joint distribution of $P(T_1, \ldots, T_m)$, where T_i is the i^{th} dimension of $\mathbf{T}$:

$$\begin{aligned} Z &\sim P_\theta(\cdot) \\ T_i|Z &\sim p_\phi(\cdot|Z) \quad i = 1, \ldots, m, \end{aligned} \tag{4.2}$$

   where Z is the latent variable. Suppose that, given $m \gg 1$, we can learn θ and ϕ such that the joint distribution of T_is from eq. (4.2) matches the observed distribution of causes $P(T_1, \ldots, T_m)$. Since T_is are conditionally independent given Z, we can show their independence structure with the graphical model in Figure 3. Note that this graph is only to reason about conditional dependencies and not a causal graph.

   Use d-separation to show that there cannot be a confounder X (other than Z) that is the parent of multiple T_is in Figure 3.

2. **[5 pts]** Given the result of the previous question, we can conclude that Z captures all the unobserved confounding. Since we can infer Z, the problem becomes a causal estimation with *observed* confounding. Describe an informal algorithm (using Back-door adjustment formula) to estimate the ATE using Z.

[1]Some examples for factor models include PCA, VAEs, ICA, LDA, etc.

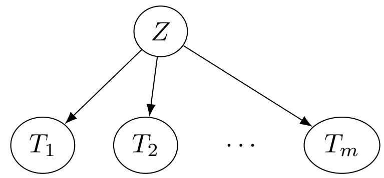


Figure 3

3. **[10 pts]** Show that the previous algorithm is wrong (!) by constructing a counterexample. In particular, show that the factorization of $\mathbf{T}$ is not unique. You can construct two SCMs $\mathcal{M}_1$, $\mathcal{M}_2$ ($X, Y \in \mathbb{R}$ and $\mathbf{T} \in \mathbb{R}^m$) with the same observed distribution $P(\mathbf{T}, Y)$ and different ATEs, where the treatments can be factorized in both, i.e.,

$$\begin{aligned} P_{\mathcal{M}_1}(T_1, \ldots, T_m | X) &= \prod_{i=1}^{m} P_{\mathcal{M}_1}(T_i | X) \\ P_{\mathcal{M}_2}(T_1, \ldots, T_m | X) &= \prod_{i=1}^{m} P_{\mathcal{M}_2}(T_i | X) \end{aligned} \tag{4.3}$$

If you cannot solve it for the general case, try constructing a numerical example for $m = 3$.

Hint: You can consider linear SCMs with Gaussian noise in the form of

$$\mathcal{M}: \quad \begin{aligned} X &:= \epsilon_X \\ \mathbf{T} &:= \alpha X + \boldsymbol{\epsilon}_{\mathbf{T}} \\ Y &:= \boldsymbol{\beta}^\top \mathbf{T} + \gamma X + \epsilon_Y \end{aligned} \tag{4.4}$$

with $\epsilon_X \sim \mathcal{N}(0, \sigma_X), \epsilon_Y \sim \mathcal{N}(0, \sigma_Y)$*, and* $\boldsymbol{\epsilon}_{\mathbf{T}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{T}})$ *(*$\boldsymbol{\Sigma}_{\mathbf{T}}$ *is a diagonal* $m \times m$ *covariance matrix). First, you need to show that* X *factorizes* $P(\mathbf{T})$ *in this SCM. Then, construct a new SCM with different coefficients and the same induced distribution* $P(\mathbf{T}, Y)$*. You can try changing the treatment assignment to*

$$\mathbf{T} := c \cdot \alpha X + \boldsymbol{\epsilon}_{\mathbf{T}} \tag{4.5}$$

and see how other coefficients should be changed.